Insensitivity of Constraint-Based Causal Discovery Algorithms to Violations of the Assumption of Multivariate Normality

Mark Voortman[*] and **Marek J. Druzdzel**[*†]

[*]Decision Systems Laboratory
School of Information Sciences and Intelligent Systems Program
University of Pittsburgh, Pittsburgh, PA 15260, USA
{voortman,marek}@sis.pitt.edu

[†]Faculty of Computer Science
Białystok Technical University
Wiejska 45A, 15-351 Białystok, Poland

Abstract

Constraint-based causal discovery algorithms, such as the PC algorithm, rely on conditional independence tests and are otherwise independent of the actual distribution of the data. In case of continuous variables, the most popular conditional independence test used in practice is the partial correlation test, applicable to variables that are multivariate Normal. Many researchers assume multivariate Normal distributions when dealing with continuous variables, based on a common wisdom that minor departures from multivariate Normality do not have a too strong effect on the reliability of partial correlation tests. We subject this common wisdom to a test in the context of causal discovery and show that partial correlation tests are indeed quite insensitive to departures from multivariate Normality. They, therefore, provide conditional independence tests that are applicable to a wide range of multivariate continuous distributions.

Introduction

Causal discovery algorithms based on constraint-based search, such as SGS and PC (Spirtes, Glymour, & Scheines 2000), perform a search for an equivalence class of causal graphs that is identifiable from patterns of conditional independencies observed in the data. These algorithms depend on the probability distribution over the variables in question only indirectly, since they take a set of conditional independence statements as input, and will correctly identify the class of causal structures that are compatible with the observed data. As long as conditional independence between random variables can be established, the algorithms produce provably correct results. The main problem to overcome is finding conditional independence tests that are suitable for a given distribution.

While reliable independence tests exist for discrete data, there are no generally accepted tests for mixtures of discrete and continuous data and even no tests that cover the general continuous case. One special case that can be tackled in practice is when the set of variables in question follows a multivariate Normal distribution. In that case, there is a well established test of conditional independence, notably partial correlation. If the partial correlation between variables X and Y conditional on a set of variables $\mathbf{Z}$ is zero, then X and Y are said to be conditionally independent. Because the multivariate Normal case is tractable, it is tempting to assume this distribution in practice. Druzdzel & Glymour (1999), for example, use a data set obtained from the US News & World Report Magazine to study causes of low student retention in US Universities. Figure 1 presents histograms of all eight variables in that data set. It seems that few, if any, of the variables are Normally distributed, so the question that naturally arises is whether they are 'Normal enough' to yield correct results when the partial correlation tests are applied. Common wisdom says that partial correlation is fairly insensitive to minor departures from Normality. What constitutes minor departures and when these departures are large enough to weaken the power of partial correlation, has, to our knowledge, not been tested systematically.

In this paper, we describe a series of experiments that we conducted to test the sensitivity of the partial correlation test, and the resulting effect on a basic causal discovery algorithm, the PC algorithm, to departures from multivariate Normality. We will show empirically that the partial correlation test is very robust against departures from Normality, and thus, in practice the PC algorithm yields rather robust results.

The PC Algorithm

The PC algorithm (Spirtes, Glymour, & Scheines 2000) makes four basic assumptions that must be satisfied to yield correct results:

1. The set of observed variables is causally sufficient. Causal sufficiency means that every common cause of two or more variables is contained in the data set. This assumption is easily relaxed (see, for example, Spirtes, Glymour, & Scheines (2000)), but this is not of much importance for the current paper.

2. All records in the data set are drawn from the same joint probability distribution.

3. The probability distribution P over the observed variables is faithful to a directed acyclic graph G of the causal structure. Faithfulness means that all and only the conditional independence relations found in P are entailed by the Markov condition applied to G. The Markov condition

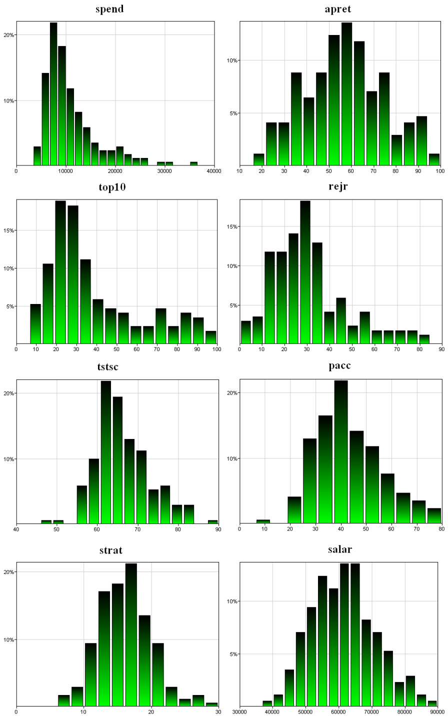


Figure 1: Marginal distributions of the 8 variables in the retention data set (Druzdzel & Glymour 1999).

is satisfied if node X in graph G is independent of all its non-descendents minus its parents, given its parents.

4. The statistical decisions required by the algorithms are correct for the population.

It is the last assumption that we focus on in this paper. The PC algorithm works as follows:

1. Start with a complete undirected graph G with vertices $\mathbf{V}$.
2. For all ordered pairs $\langle X, Y \rangle$ that are adjacent in G, test if they are conditionally independent given a subset of $\mathbf{Adjacencies}(G, X) \setminus \{Y\}$. We increase the cardinality of the subsets incrementally, starting with the empty set. If the conditional independence test is positive, we remove the undirected link and set $\mathbf{Sepset}(X, Y)$ and $\mathbf{Sepset}(Y, X)$ to the conditioning variables that made X and Y conditionally independent.
3. For each triple of vertices X, Y, Z, such that the pairs $\{X, Y\}$ and $\{Y, Z\}$ are adjacent in G but $\{X, Z\}$ is not, orient $X - Y - Z$ as $X \rightarrow Y \leftarrow Z$ if and only if Y is not in $\mathbf{Sepset}(X, Z)$.
4. Orient the remaining edges in such a way that no new conditional independencies and no cycles are introduced. If an edge could still be directed in two ways, leave it undirected.

We illustrate the PC algorithm by means of a simple example (after Druzdzel & Glymour (1999)). Suppose we obtained a data set that is generated by the causal structure in Figure 2a, and we want to rediscover this causal structure.

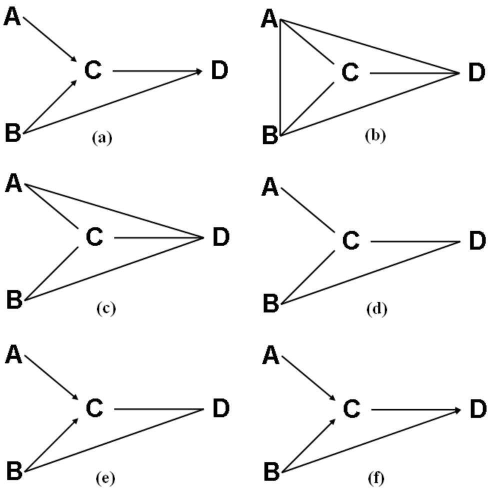


Figure 2: (a) The underlying directed acyclic graph. (b) The complete undirected graph. (c) Graph with zero order conditional independencies removed. (d) Graph with second order conditional independencies removed. (e) The partially rediscovered graph. (f) The fully rediscovered graph.

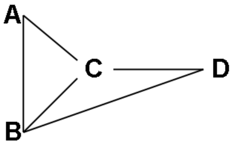


Figure 3: Resulting graph when the statistical test failed to find the independency $A \perp B$.

In Step (1), we start out with a complete undirected graph, shown in Figure 2b. In Step (2) we remove an edge when two variables are conditionally independent on a subset of adjacent variables. The graph in Figure 2 implies two (conditional) independencies (denoted by $\perp$), namely $A \perp B$ and $A \perp D | \{B, C\}$, which leads to graphs in Figure 2c and 2d, respectively. Step (3) is crucial, since it is in this step where we orient the causal arcs. In our example, we have the triplet $A - C - B$ and C is not in $\mathbf{Sepset}(A, B)$, so we orient $A \rightarrow C$ and $B \rightarrow C$ in Figure 2e. In Step (4) we have to orient $C \rightarrow D$, otherwise $A \perp D | \{B, C\}$ would not hold, and $B \rightarrow D$ to prevent a cycle. Figure 2(f) shows the final result. In this example, we are able to rediscover the complete causal structure, although this is not possible in general.

It is important to note the impact of an incorrect statistical decision. If, for example, our statistical test failed to find $A \perp B$, the resulting graph would be that of Figure 3. In general, failing to find just one conditional independence could have a severe impact on the output of the PC algorithm.

Partial Correlation Test

If P is a probability distribution linearly faithful to a graph G, then A and B are conditionally independent given $\mathbf{C}$ if and only if the partial correlation is zero, i.e., $\rho_{AB.\mathbf{C}} = 0$. Partial correlation $\rho_{AB.\mathbf{C}}$ is the correlation between residuals R_1 and R_2 resulting from the linear regression of A on $\mathbf{C}$ and B on $\mathbf{C}$, respectively.

Since the sample partial correlation will almost never be exactly zero, we have to resort to a statistical test to make a decision and this is where the assumption of multivariate Normality shows up. Fisher's z-transform (Fisher 1915) assumes a multivariate Normal distribution and is given by:

$$z(\hat{\rho}_{AB.\mathbf{C}}) = \frac{1}{2}\sqrt{n - |\mathbf{C}| - 3}\log\left(\frac{1+\hat{\rho}_{AB.\mathbf{C}}}{1-\hat{\rho}_{AB.\mathbf{C}}}\right) ,$$

where $\hat{\rho}_{AB.\mathbf{C}}$ is the sample partial correlation, n is the number of samples, and $|\mathbf{C}|$ denotes the number of variables in $\mathbf{C}$. Because $z(\hat{\rho}_{AB.\mathbf{C}}, n)$ will follow the standard Normal distribution, we can apply the z-test with $H_0 : \hat{\rho}_{AB.\mathbf{C}} = 0$ and $H_1 : \hat{\rho}_{AB.\mathbf{C}} \neq 0$. We reject H_0 at a significance level α if

$$|z(\hat{\rho}_{AB.\mathbf{C}})| > \Phi^{-1}(1-\alpha/2) ,$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function of the standard Normal distribution, also known as the probit function.

One sufficient condition for a distribution to be multivariate Normal is the following. A random vector $X = [X_1, \ldots, X_n]^T$ follows a multivariate Normal distribution if every linear combination $Y = a_1X_1 + \ldots + a_NX_N$ is Normally distributed. Because a sum of linearly dependent Normal variables is Normally distributed, a sufficient condition for Y to be multivariate Normal is that the marginal distributions of the variables X_i in the data set are Normal and that the variables depend on each other only linearly.

Experiments

In this section, we show a series of experiments that aim at testing how sensitive the partial correlation test is to violations of the assumption of multivariate Normality. The setup of all the experiments is similar and consists of three steps:

1. Generate a data set of 100 records from a joint probability distribution (the distribution is different in every experiment).
2. For each data set, perform a series of partial correlation tests, with $\alpha = 0.05$. For simplicity, we introduce only one conditioning variable in each test, unless explicitly stated otherwise. The null hypothesis is that the given variables are conditionally independent, and the alternative hypothesis is that they are not.
3. Repeat Steps (1) and (2) 1,000 times.

There are many ways for a distribution to be non-Normal and, therefore, finding out the effect of non-Normal distributions on partial correlation tests is somewhat like probing in the dark. To make this investigation systematical, we focussed on the third and fourth central moments around the mean, skewness, and kurtosis, respectively. Because both central moments are zero for the Normal distribution, and most of the time non-zero for other distributions, this is one measure of closeness to Normality.

In the first experiment, we focus on the third central moment using the Pearson IV distribution, because it is capable of keeping the kurtosis constant and change the skewness. The second experiment is the converse, namely fixing the skewness and changing the kurtosis, using the Pearson VII distribution. The third experiment changes both of the moments at the same time using the Gamma distribution. As a last experiment, we investigate the effect of multiple conditioning variables on the partial correlation test.

In the experiments, we measure the effect of using non-Normal distribution by counting Type I and Type II errors. A Type I error is committed when the null hypothesis is falsely rejected, and, conversely, a Type II error is committed when the null hypothesis is falsely accepted. The value for α is the probability that the statistical test will make a Type I error. The probability of a Type II error depends on α, the sample size, and the sensitivity of the data.

Manipulation of the Third Central Moment

We used the Pearson type IV distribution (Heinrich 2004) to manipulate the skewness while keeping the kurtosis constant. The Pearson type IV distribution is given by:

$$f(x|m,\nu) = k\left(1+x^2\right)^{-m}\exp(-\nu\tan^{-1}(x)) ,$$

where $m > 1/2$ and k is a normalization constant. We chose standard values for the location and scale parameters, because they do not influence the skewness and kurtosis. Skewness γ_1 and kurtosis γ_2 are given by:

$$\begin{array}{rcl} \gamma_1 & = & \frac{-4\nu}{r-2}\sqrt{\frac{r-1}{r^2+\nu^2}} \\ \gamma_2 & = & \frac{3(r-1)[(r+6)(r^2+\nu^2)-8r^2]}{(r-2)(r-3)(r^2+\nu^2)} \end{array} ,$$

where $r = 2(m-1)$ and $m > 5/2$ for the kurtosis to exist. The data were generated by the following joint distributions:

$$\begin{cases} X & \sim & \text{PearsonIV}(r,\nu) \\ Y & \sim & X + \text{PearsonIV}(r,\nu) \\ Z & \sim & X + \text{PearsonIV}(r,\nu) \end{cases}$$

$$\begin{cases} X & \sim & \text{PearsonIV}(r,\nu) \\ Y & \sim & \text{PearsonIV}(r,\nu) \\ Z & \sim & X + Y + \text{PearsonIV}(r,\nu) \end{cases}$$

Graphs implied by the distributions used in the experiments are depicted in Figure 4a and 4b, respectively.

We used the following values for r and ν such that the kurtosis is kept constant:

r	ν	Skewness	Kurtosis
4	0	0	9
5	2.40	-1.15	9
6	4.90	-1.41	9
7	9.04	-1.55	9
8	19.60	-1.63	9

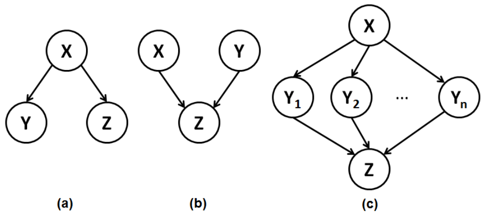


Figure 4: Graphs implied by the distributions used in the experiments. (a) Graph with a common cause. (b) Graph with a common effect. (c) Graph with multiple conditioning variables.

The skewness is bounded by the value for kurtosis and the range of allowable parameter values in the Pearson IV family and, therefore, we could only achieve a skewness of -1.63. For all generated data sets, we tested the correlation and partial correlation between all $\binom{3}{2} \cdot 2 = 6$ possible combinations of variables. In this setting, the partial correlation test performed as expected without any surprising results. The following are the results for the graph in Figure 4a:

Skew	XY	XZ	YZ	$XY\|Z$	$XZ\|Y$	$YZ\|X$
0	0	0	0	0.002	0	0.043
-1.15	0	0	0.003	0	0	0.052
-1.41	0	0	0.001	0.001	0	0.061
-1.55	0	0	0.001	0.001	0	0.047
-1.63	0	0	0.003	0.001	0	0.044

The numbers in the table denote the error frequencies. So, for example, when the skewness is -1.55, the error frequency of the partial correlation of Y and Z given X is equal to 0.047. This means that about 5 percent of the time the partial correlation test lead to an incorrect result, which is exactly what we would expect with $\alpha = 0.05$. Also, when the variables were (conditionally) dependent, which is the case for all the other combinations of variables, the test made very few mistakes.

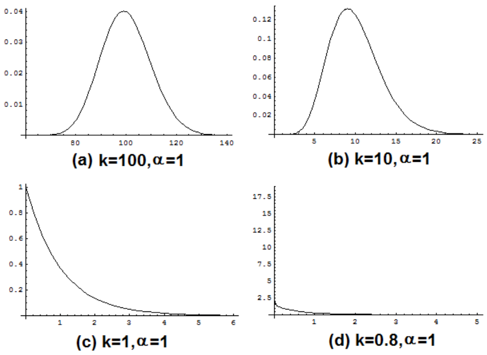
Similar results were obtained for the graph in Figure 4b:

Skew	XY	XZ	YZ	$XY\|Z$	$XZ\|Y$	$YZ\|X$
0	0.054	0	0	0	0.005	0
-1.15	0.045	0	0	0.002	0	0
-1.41	0.042	0	0	0.004	0	0
-1.55	0.068	0	0	0	0.004	0
-1.63	0.054	0	0	0	0.005	0

Manipulation of the Fourth Central Moment

In the second experiment, we investigated the impact of changing the kurtosis of a distribution, while keeping all other moments constant. To achieve this goal, we used the Pearson type VII distribution of which the density function is given by:

$$f(x|a, m) = \frac{\Gamma(m)}{a\sqrt{\pi}\Gamma(m - 1/2)} \left[1 + \left(\frac{x}{a}\right)^2\right]^{-m} ,$$



(a) k=100,α=1

(b) k=10,α=1

(c) k=1,α=1

(d) k=0.8,α=1

Figure 5: Gamma distributions for various values of the shape parameter k. Please note that for $k < 1$ the distribution becomes extremely skewed.

where a is a scale parameter and m is a shape parameter. We reparametrize with $a = \sqrt{2 + 6/\gamma_2}$ and $m = 5/2 + 3/\gamma_2$, where γ_2 is the kurtosis. The data were generated by the following joint distributions:

$$\begin{cases} X & \sim & \text{PearsonVII}(\gamma_2) \\ Y & \sim & X + \text{PearsonVII}(\gamma_2) \\ Z & \sim & X + \text{PearsonVII}(\gamma_2) \end{cases}$$

$$\begin{cases} X & \sim & \text{PearsonVII}(\gamma_2) \\ Y & \sim & \text{PearsonVII}(\gamma_2) \\ Z & \sim & X + Y + \text{PearsonVII}(\gamma_2) \end{cases}$$

The Pearson VII family of distributions is symmetric, so the skewness is always zero. γ_2 ranged from 1 to 20 with increments of 1. The partial correlation test performed similar to the results in the previous experiment, i.e., it did not affect the probability of a Type I error and made very few Type II errors.

Manipulation of the Gamma Distribution

In the third experiment, we changed the skewness and kurtosis at the same time. For this we used the Gamma distribution, given by:

$$f(x|k, \theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} e^{-x/\theta} ,$$

where $k > 0$ is a shape parameter and $\theta > 0$ is a scale parameter. The skewness is given by $\gamma_1 = 2/\sqrt{k}$ and the kurtosis is given by $\gamma_2 = 6/k$. One important property of the Gamma distribution is that it approximates the Normal distribution when k goes to infinity. Therefore, we start our experiments with large k and decrease its value to see the effect on the statistical tests. Figure 5 shows four different Gamma distributions that change from approximately Normal to very non-Normal distributions. Note that decreasing k increases the skewness and kurtosis. We performed this experiment in a similar setup as the previous experiments. The data were generated by using the following system of simultaneous equations:

$$\begin{cases} X & \sim & \text{Gamma}(k, 1) \\ Y & \sim & X + \text{Gamma}(k, 1) \\ Z & \sim & X + \text{Gamma}(k, 1) \end{cases}$$

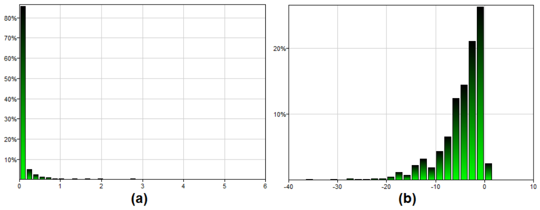


Figure 6: (a) Histogram of 1,000 samples from a Gamma distribution with $k = 0.1$ and $\theta = 1$. (b) Histogram of the logarithm of the samples in (a).

$$\begin{cases} X & \sim & \mathrm{Gamma}(k, 1) \\ Y & \sim & \mathrm{Gamma}(k, 1) \\ Z & \sim & X + Y + \mathrm{Gamma}(k, 1) \end{cases}$$

Figure 4a and 4b show the causal structures implied by the above systems of equations. Please note that the marginal distributions of X, Y, and Z will be Gamma distributions, because a sum of independent random variables following Gamma distributions with the same scale parameter will follow a Gamma distribution.

For k we chose the values 1,000, 100, 10, 1, 0.1, and 0.01. For these values, we have the following skewness and kurtosis:

k	Skew	Kurt
1,000	0.063	0.006
100	0.2	0.06
10	0.63	0.6
1	2	6
0.1	6.32	60
0.01	20	600

Please note that setting very high theoretical values will not result in limited size samples that match these values, but nonetheless they are relatively higher than samples with a lower theoretically chosen skewness and kurtosis. Our experiments support this.

The partial correlation test turned out to be very robust in most of these cases. It made a Type I error around 5 percent of the time as one would expect, and when $k < 0.1$, this percentage even approached zero. When we consider Type II errors, the test makes very few errors in most of the cases. Only if $k < 0.1$, the test increases its frequency of Type II errors, but the number of errors never exceeds 35 percent. Also, please note that $k < 0.1$ yields a distribution that is extremely skewed ($\gamma_1 > 6.32$). Figure 6 shows a histogram of data generated from a $\mathrm{Gamma}(0.1, 1)$ distribution, where it is clear that the distribution is skewed to the right. Figure 6 shows the histogram of the logarithm of the samples.

We also ran several trials, in which we randomly selected different values for k. In this case, the tests again were very robust. Only when at least one of the k's was close to zero the tests yielded incorrect results.

Multiple Conditioning Variables

As a last experiment, we investigated the effect of having multiple, instead of only one, conditioning variables. Statistical test do not work quite as well compared to the situation with only one conditioning variable. In this experiment, we wanted to find out if this effect on a non-Normal distribution is different than on a Normal distribution. We generated the data by sampling from the following joint probability distribution:

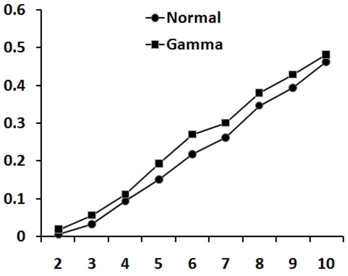


Figure 7: Results for the network with multiple conditioning variables. The horizontal axis shows the number of conditioning variables, and the vertical axis shows the error frequency. Clearly, a smaller number of conditioning variables yields more accurate results.

$$\begin{cases} X & \sim & \mathrm{Gamma}(k_X, 1) \\ Y_1 & \sim & X + \mathrm{Gamma}(k_1, 1) \\ Y_2 & \sim & X + \mathrm{Gamma}(k_2, 1) \\ \vdots & & \\ Y_n & \sim & X + \mathrm{Gamma}(k_n, 1) \\ Z & \sim & Y_1 + Y_2 + \ldots + Y_n + \mathrm{Gamma}(k_Z, 1) \end{cases}$$

All the k_i were drawn from a $\mathrm{Uniform}(0, 1)$ distribution. The graph implied by this joint distribution is depicted in Figure 4c.

We calculated the partial correlation between X and Z given $\{Y_1, \ldots, Y_{n-1}\}$, where we ranged n from 3 to 10. For comparison, we ran another trial with the same setup, but replaced the Gamma distributions by standard Normal distributions. The results are displayed in Figure 7. Although the setting with the Gamma distribution performs not as well as the standard Normal case, the difference is very small. Especially when we take into account the fact that the Gamma distributions used in the experiments had $k_i \sim \mathrm{Uniform}(0, 1)$ and are all extremely skewed to the right (see Figure 5), the result is surprising. Please note that 100 samples is quite a small number for so many conditioning variables. We ran the same experiment with 1,000 samples, and this resulted in no errors for both the Normal and Gamma distributions.

Violations of Linearity

In the previous section, we showed that partial correlation tests are quite robust against departures from the multivariate Normality assumption, although we have not tested the influence of departures from linearity of interactions. It is well known that non-linearity can make the partial correlation test fail. To get an idea how stong the non-linearity has to be for the partial correlation test to fail, we created the

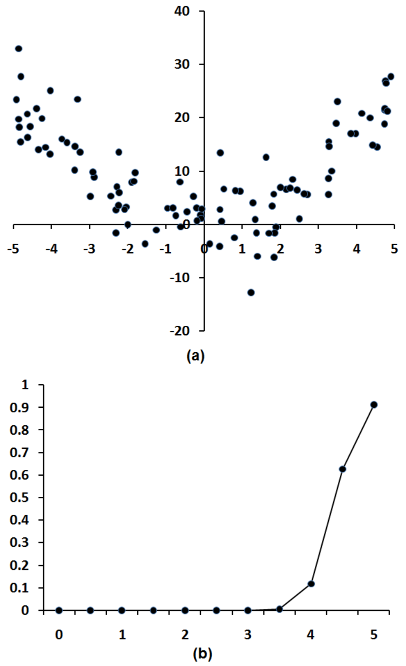


Figure 8: (a) Scatterplot of one of the samples when $a = 5$. (b) a on the horizontal axis, versus the error frequency on the vertical axis.

following example:

$$\begin{cases} X & \sim & \text{Uniform}(-5, a) \\ Y & \sim & X^2 + \text{Normal}(0, 5) \end{cases}$$

The idea here is that changing a has an effect on how non-linear the relation beween X and Y is. Increasing a will have the effect that the dependence of Y on X will change from almost a linear to a quadratic dependency on X. We sampled 100 data points for every a that ranges from 0 to 5 with increments of 0.5, and performed correlation tests for each of these cases. We display the results in Figure 8. As you can see, the correlation tests only failed when a approached 5. The underlying reason is that the regression line becomes almost horizontal, so knowing X gives almost no information about the value of Y. Obviously, non-linearity of interactions can have a large influence on the tests and in this way the tests can be fooled.

Discussion

We have shown experimentally that the partial correlation test is quite robust against deviations from multivariate Normal distributions. All of our experiments support this claim. Only if the deviation from multivariate normality is very large, the tests might yield incorrect results. This means that assuming a multivariate Normal distribution in practice is reasonable and one could look at the histograms to see if the distributions are not too non-Normal. Note that we only used a sample size of 100, and testing with a sample size of 1,000 gave only slightly better results. It is likely that there are cases not covered by us, where the partial correlation is quite sensitive to the distribution in use, e.g., a multimodal distribution. The rule of thumb seems to be that as long as the distributions are not too non-Normal, the statistical tests will lead to reasonable results and so will the causal discovery algorithms. The problems of non-linear relationships in the data is harder to deal with and we presented merely a simple example that visually shows a case when the test will go wrong.

There are two lines of work that take an alternative approach by not assuming multivariate Normal distributions at all. In Shimizu *et al.* (2005), the opposite assumption is made, namely that all error terms are non-Normally distributed. This allows them to find the complete causal structure, while also assuming linearity and causal sufficiency, something that is not possible for Normal error terms. Of course, this brings up the emperical question whether error terms are typically distributed Normally or non-Normally.

The second approach does not make any distributional assumptions at all. Margaritis (2005) describes an approach that is able to perform conditional independence tests on data that can have any distribution. However, the practical applicability of the algorithm is still an open question.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and Peter Spirtes for sharing with us his views on the problem and for useful literature pointers. This research was supported by the Air Force Office of Scientific Research grant FA9550–06–1–0243 and by Intel Research. All experimental data have been obtained using SMILE, a Bayesian inference engine developed at the Decision Systems Laboratory and available at `http://genie.sis.pitt.edu/`.

References

Druzdzel, M. J., and Glymour, C. 1999. Causal inferences from databases: Why universities lose students. In Glymour, C., and Cooper, G. F., eds., *Computation, Causation, and Discovery*, 521–539. Menlo Park, CA: AAAI Press.

Fisher, R. 1915. Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* 507–521.

Heinrich, J. 2004. A guide to the Pearson Type IV distribution. Cdf/memo/statistics/public/6820.

Margaritis, D. 2005. Distribution-free learning of Bayesian network structure in continuous domains. In *Proceedings of The Twentieth National Conference on Artificial Intelligence (AAAI)*.

Shimizu, S.; Hyvarinen, A.; Kano, Y.; and Hoyer, P. O. 2005. Discovery of non-Gaussian linear causal models using ICA. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, 525–53. Arlington, Virginia: AUAI Press.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search.* New York: Springer Verlag, second edition.